Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 67-74 recibido 29-03-2016 revisado 21-04-2016 aceptado 02-05-2016

Una aproximación al uso de *word embeddings* en una tarea de similitud de textos en español

An approach to the use of word embeddings in a textual similarity task for Spanish texts

Tomás López-Solaz Universidad de Sevilla tlopez2@us.es
José A. Troyano Universidad de Sevilla troyano@us.es
F. Javier Ortega Universidad de Sevilla javierortega@us.es
Fernando Enríquez Universidad de Sevilla fenros@us.es

Resumen: En este trabajo mostramos cómo una representación vectorial de palabras basada en *word embeddings* puede ayudar a mejorar los resultados en una tarea de similitud semántica de textos. Para ello hemos experimentado con dos métodos que se apoyan en la representación vectorial de palabras para calcular el grado de similitud de dos textos, uno basado en la agregación de vectores y otro basado en el cálculo de alineamientos. El método de alineamiento se apoya en la similitud de vectores de palabras para determinar la vinculación entre las mismas. El método de agregación nos permite construir representaciones vectoriales de los textos a partir de los vectores individuales de palabras. Estas representaciones son comparadas mediante dos distancias clásicas como son la euclídea y la del coseno. Hemos evaluado nuestros sistemas con el corpus basado en Wikipedia distribuido en la competición de similitud de textos en español de SemEval-2015. Nuestros experimentos muestran que el método basado en alineamiento se comporta mucho mejor, obteniendo resultados muy cercanos al mejor sistema de SemEval. El método basado en agregación de vectores se comporta sensiblemente peor. No obstante, esta segunda aproximación parece capturar aspectos de similitud no recogidos por la primera, ya que cuando se combinan las salidas de ambos sistemas se mejoran los resultados del método de alineamiento, superando incluso los resultados del mejor sistema de SemEval.
Palabras clave: Similitud semántica, *word embedding*, alineamiento de textos

Abstract: In this paper we show how a vector representation of words based on word embeddings can help to improve the results in tasks focused on the semantic similarity of texts. Thus we have experimented with two methods that rely on the vector representation of words to calculate the degree of similarity of two texts, one based on the aggregation of vectors and the other one based on the calculation of alignments. The alignment method relies on the similarity of word vectors to determine the semantic link between them. The aggregation method allows us to construct vector representations of the texts from the individual vectors of each word. These representations are compared by means of two classic distance measures: Euclidean distance and cosine similarity. We have evaluated our systems with the corpus based on Wikipedia distributed in the competition of similarity of texts in Spanish of SemEval-2015. Our experiments show that the method based on the alignment of words performs much better, obtaining results that are very close to the best system at SemEval. The method based on vector representations of texts behaves substantially worse. However, this second approach seems to capture aspects of similarity not detected by the first one, as when the outputs of both systems are combined the results of the alignment method are surpassed, even exceeding the results of the best system at SemEval.
Keywords: Semantic similarity, *word embedding*, text alignment

ISSN 1135-5948

1 Introducción

La medición fiable de la similitud de textos es una aplicación de gran ayuda para distintas tareas relacionadas con el procesamiento de textos en lenguaje natural. Sólo por citar algunos ejemplos, los sistemas de clasificación de documentos, de resúmenes de textos, o de traducción automática, pueden beneficiarse de un módulo de similitud que establezca el grado de parecido entre dos unidades textuales. En general, podemos distinguir entre dos tipos de similitud: a nivel de palabras y a nivel de textos.

Las distintas aproximaciones de similitud a nivel de palabras se agrupan en dos categorías: similitud léxica de palabras y similitud semántica de palabras. Dos palabras son similares a nivel léxico si están compuestas por secuencias parecidas de caracteres. Para determinar la similitud léxica de palabras se suelen utilizar distintas métricas basadas en comparación de cadenas de caracteres. La similitud semántica de palabras, por su parte, nos permite medir si dos palabras tienen significados parecidos o se usan en contextos parecidos. Hay dos grandes grupos de técnicas para calcular la similitud semántica de palabras: técnicas basadas en conocimiento y técnicas basadas en corpus.

Las técnicas de similitud de palabras basadas en conocimiento miden el grado de parecido entre palabras apoyándose en algún tipo de recurso lingüístico que proporcione información sobre el significado de las palabras. El recurso por excelencia es WordNet (Miller, 1995), una base de datos léxica organizada en torno a varias relaciones entre palabras. La relación de sinonimia es la más importante y en ella se apoya el concepto de *synset* (grupo de sinónimos) que permite definir de forma implícita el significado de las palabras a través del conjunto de *synsets* en los que aparece. En función de las relaciones entre palabras, se han definido varias métricas de similitud entre las que destacan las de Resnik (Resnik, 1995), Lin (Lin, 1998) y Jian & Conrath (Jiang y Conrath, 1997).

Las técnicas de similitud de palabras basadas en corpus determinan el parecido semántico de dos palabras en función de los usos de esas palabras en una gran colección de textos. Por lo general, estas técnicas se basan en algún tipo de representación vectorial de las palabras en función de los distintos contextos en los que dichas palabras aparecen. Dentro de las técnicas basadas en corpus destacan las de *latent semantic analysis* (LSA) y las de *word embedding*. LSA analiza las relaciones entre un conjunto de documentos y los términos que contienen, estableciendo que dos palabras son similares si ocurren en fragmentos similares de textos. LSA parte de una matriz de palabras frente a documentos y aplica una técnica matemática denominada *singular value decomposition* que permite reducir el número de filas (documentos) preservando la similitud entre columnas (palabras). Por su parte, las técnicas de *word embedding* parten de representaciones BOW (*bag of words*) de los distintos contextos de las palabras para obtener representaciones vectoriales de las palabras de dimensiones mucho más reducidas que capturan el significado y las relaciones entre palabras. Hay diversas técnicas para calcular estas representaciones, una de las más empleadas se basa en redes neuronales de una sola capa oculta que predicen la palabra dado el contexto o viceversa, adaptando así una de las piezas básicas de los modelos de aprendizaje profundo, los autocodificadores. Las técnicas de *word embedding* han demostrado ser muy útiles en múltiples tareas del Procesamiento del Lenguaje Natural aparte de la similitud de textos (Collobert et al., 2011), (Zou et al., 2013), y en la actualidad gozan de gran popularidad. Esto se debe en gran medida a la existencia de herramientas como Word2vec (Mikolov et al., 2013) o GloVe (Pennington, Socher, y Manning, 2014) que han facilitado mucho el acceso a este tipo de técnicas a la comunidad investigadora.

Las técnicas de similitud a nivel de palabras proporcionan una información básica para afrontar la tarea de similitud a nivel de textos. Muchas de las aproximaciones de similitud de textos se basan en la idea de alineamiento, que básicamente consiste en un emparejamiento entre palabras de los dos textos a comparar.

El alineamiento entre dos textos proporciona un marco sobre el que se pueden evaluar distintas heurísticas para determinar el grado de similitud entre los textos. Por ejemplo, usando las conexiones propuestas en el alineamiento para calcular métricas de similitud semántica entre las palabras emparejadas y agregando estas métricas individuales para obtener una métrica de similitud global entre los dos textos.

En este trabajo estamos interesados en analizar vías que permitan integrar el conocimiento obtenido mediante *word embeddings* a la hora de determinar el grado de similitud de dos textos. El uso de *word embeddings* no es nuevo en sistemas de similitud de textos, pero siempre como complemento a otras técnicas o recursos. Nuestra intención es evaluar el comportamiento de un sistema que se base exclusivamente en el conocimiento proporcionado por un modelo de *word embedding*.

Hemos identificado dos maneras en las que la representación de palabras en el espacio continuo puede ser de utilidad para esta tarea: con un método de alineamiento basado exclusivamente en *embeddings* y con una representación vectorial de textos basada en los vectores de palabras.

El método de alineamiento propuesto usa el grado de similitud de palabras como mecanismo para identificar posibles emparejamientos de palabras, construyéndose alineamientos bidireccionales en los que las palabras son emparejadas parcialmente en función de su similitud.

La segunda aproximación consiste en el uso de los vectores de palabras para construir representaciones vectoriales de los textos. Estas representaciones son comparadas mediante dos distancias clásicas como son la euclídea y la del coseno para obtener así un grado de similitud entre dos textos.

Hemos evaluado nuestros sistemas con el corpus basado en wikipedia distribuido en la competición de similitud de textos en español de SemEval-2015. De nuestras dos aproximaciones, la que mejor se comporta es la del método de alineamiento, para la que se obtienen resultados muy cercanos al mejor sistema de SemEval. La idea de usar los vectores de palabras para construir vectores para los textos se comporta sensiblemente peor. No obstante esta segunda aproximación parece capturar aspectos de similitud no recogidos por la primera, ya que cuando se combinan las salidas de ambos sistemas se mejoran los resultados del método de alineamiento, superando incluso los resultados del mejor sistema de SemEval.

El resto del artículo se organiza de la siguiente forma: la sección 2 describe brevemente algunos trabajos relacionados, la sección 3 describe tanto el método basado en representación vectorial de textos como el basado en alineamiento, así como el método de combinación aplicado, la sección 4 incluye los resultados experimentales y, por último, la sección 5 incluye las conclusiones y plantea algunas líneas de trabajo futuro.

2 Trabajos relacionados

La mayoría de las contribuciones recientes en el ámbito de la similitud semántica de textos provienen de los participantes en las tareas que se proponen anualmente en SemEval[1]. De estas participaciones han surgido muchas técnicas, ideas, e incluso frameworks que ofrecen componentes que pueden ser utilizados para desarrollar sistemas de similitud de textos, como por ejemplo DKPro (Bär, Zesch, y Gurevych, 2013). Aparte de los trabajos presentados en SemEval, otro referente clásico en este campo es el trabajo de (Mihalcea, Corley, y Strapparava, 2006) centrado en el cálculo de similitud semántica para textos cortos, y en el que se resumen las métricas de similitud más importantes que posteriormente se han venido utilizando como parte de la mayoría de los sistemas de similitud de textos.

Las distintas propuestas que han participado en las ediciones recientes de las tareas de similitud de SemEval son combinaciones de métricas clásicas con ideas originales. En muchos casos, los participantes van mejorando sus sistemas año a año para incorporar nuevas ideas o para adaptarse a los distintos cambios en la definición de las tareas por parte de los organizadores. A modo de muestra del tipo de técnicas usadas en estos sistemas, resumimos a continuación las características más significativas de los tres mejores sistemas en la tarea en español de la competición de SemEval de 2015 (Agirre et al., 2015), cuyo corpus hemos usado en nuestros experimentos.

El mejor grupo fue (Hänig, Remus, y Puente, 2015), su solución integra tres técnicas: representación vectorial de textos (BOW y distancia del coseno), alineamiento mediante métricas de similitud y uso de *machine learning* para la combinación de las distintas métricas calculadas. Usa un alineamiento secuencial y emplea Word2vec en una última fase para intentar emparejar palabras residuales que no han sido emparejadas por su técnica de alineamiento.

El segundo grupo de la competición (Ka-

[1] https://en.wikipedia.org/wiki/SemEval

rumuri, Vuggumudi, y Chitirala, 2015) se apoya en un sistema previo para inglés y utiliza el traductor de Google para adaptar las entradas al español a su sistema. Integra un sistema de alineamiento con distintas métricas basadas en proporcionalidad, número de sustantivos, número de adjetivos, tamaño de los textos, etc.

El tercer grupo de la competición (Biçici, 2015) presentó un sistema basado en máquinas de traducción referencial. La idea principal es comparar cómo se parecen los textos originales cuando son traducidos a otros idiomas.

3 Método propuesto

El método propuesto para el cálculo de la similitud entre textos se basa en la combinación de diversos indicadores en los que el factor común es el uso de *word embeddings* para representar las palabras.

3.1 Agregando *word embeddings*

La primera aproximación que forma parte de nuestro sistema consiste en la aplicación de alguna medida de similitud entre los vectores que proporciona la representación de *word embeddings*.

Dado que estos vectores están asociados a palabras individuales, para cada texto es necesario aplicar algún método que unifique los vectores de cada palabra generando un único vector. Tras probar distintas funciones de agregación, se eligió como representación la media aritmética de los vectores, obtenida sumando todos los vectores de palabras y dividiendo entre el número de palabras que componen el texto (ecuación 1).

$$\vec{V_d} = \frac{\sum_{i=0}^{n} \vec{v_i}}{n} \quad (1)$$

Una vez obtenido el vector agregado para cada texto se aplican medidas de similitud 'tradicionales', como son la distancia euclídea y la similitud del coseno.

3.2 Alineamiento

Otro indicador que obtenemos sobre el grado de similitud entre los textos se basa en la idea general del alineamiento de palabras. En este caso se lleva a cabo el emparejamiento de las palabras que, perteneciendo cada una a un texto distinto, guardan alguna relación semántica entre ellas. Una vez completada la fase de alineamiento de palabras se utiliza la distancia entre los *word embeddings* de cada palabra para finalmente aplicar una función de agregación que nos proporcione el grado de similitud global entre los textos.

El alineamiento se realiza habitualmente en un único sentido, buscando para cada palabra de la oración de menor tamaño aquella que debe formar pareja con ella de entre todas las que forman la oración más larga. Sin embargo, en nuestro método realizamos un alineamiento bidireccional, tras el cual se descartan los pares de palabras repetidos antes de tomar la decisión final. De este modo contamos con un mayor número de medidas parciales mejorando los resultados finales.

El proceso comienza con la construcción de un vector en el que existe una posición asociada a cada palabra del conjunto formado por la unión de los dos textos a analizar. Para cada palabra del primer texto se buscará una palabra del segundo texto con la que alinearla. Si la misma palabra existe en los dos textos se alineará consigo misma, y en caso contrario se calculará la distancia entre las representaciones vectoriales de la palabra de referencia del primer texto y todas las del segundo, con el fin de buscar la más cercana. El valor a introducir en la posición del vector asociada a la palabra de referencia será un 1 en el primer caso, y 1 menos la distancia en el segundo. Si el alineamiento no se puede llevar a cabo se introducirá un 0. Esto puede suceder si la palabra no está presente en el vocabulario del modelo de *word embeddings* utilizado.

El vector resultante representa indicadores individuales de similitud para las palabras que aparecen en los textos, por lo que es necesario aplicar una función de agregación.

En el algoritmo 1 se describe el proceso de forma más detallada, siendo posible aplicar distintas configuraciones para obtener las *selected_words* que se utilizan para generar el valor agregado de similitud (número total de palabras, palabras con valor distinto de cero, etc.).

En la figura 1 se muestra un ejemplo de alineamiento entre dos textos, incluyendo los valores de similitud obtenidos para cada par de palabras. El vector resultante tendrá once posiciones (el número de palabras distintas presentes en ambos textos) y el grado de similitud se calcula en este caso con las posiciones cuyo valor es distinto de cero, siendo el resultado $0,718$ sobre 1. La figura refleja la capa-

4	Dos oraciones son completamente equivalentes al significar la misma cosa.
3	Dos oraciones son prácticamente equivalentes pero algunos detalles difieren.
2	Dos oraciones son aproximadamente equivalentes pero alguna información importante difiere o no está.
1	Dos oraciones no son equivalentes pero son del mismo tema.
0	Dos oraciones son de diferentes temas.

Tabla 1: Descripción de los niveles de similitud para la tarea de SemEval 2015

Algoritmo 1 Alineamiento en pseudocódigo

Require: t_1, t_2 sets de tokens, m = modelo Word2Vec
Ensure: sim = similitud métrica

$vocab = t_1 + t_2$
$bow = [0] * vocab.size()$
for all w **in** t_1 **do**
  if w **in** t_2 **then**
    $bow[w] = 1$
  else if w **in** m **then**
    $bow[w] = max(m.similarity(w, t_2))$
  else
    $continue$
  end if
end for
for all w **in** t_2 **do**
  if w **in** m **then**
    $bow[w] = max(m.similarity(w, t_1))$
  else
    $continue$
  end if
end for
$values = (w\ for\ w\ in\ selected\_words)$
$sim = sum(values)/values.size()$
return sim

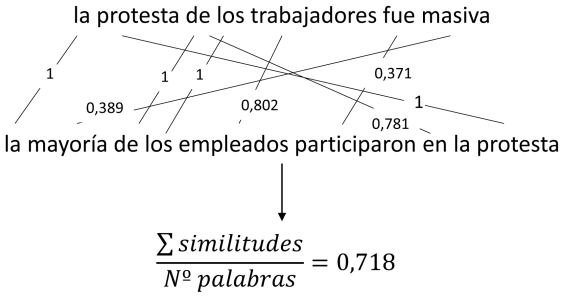

Figura 1: Ejemplo de alineamiento

cidad de establecer relaciones semánticas entre palabras distintas a través del modelo de *word embedding* ('trabajadores'-'empleados') y la posibilidad de tener una palabra alineada con más de una palabra distinta ('de').

3.3 Combinación

Con el objetivo de aprovechar los diferentes enfoques aportados por las soluciones propuestas en las secciones anteriores, hemos experimentado con un método de combinación basado en un algoritmo de aprendizaje automático supervisado para regresión. En este caso se construye un conjunto de datos de entrenamiento formado por los valores devueltos por los sistemas anteriormente descritos, generando un nuevo modelo que se pueda aplicar a nuevos pares de textos que se deseen analizar.

4 Experimentación

Como marco de referencia para evaluar el rendimiento de nuestra propuesta se recurrió a la última edición de la conferencia internacional SemEval, celebrada en 2015 (Agirre et al., 2015).

Entre los diferentes retos que se presentaron en este evento, hemos seleccionado la tarea para el español que consiste en, dadas dos oraciones, devolver un valor de similitud continuo entre 0 y 4. Para entender la diferencia de matices que hay entre los distintos valores, mostramos en la tabla 1 una pequeña descripción de cada uno de los cinco niveles en que se ha dividido el grado de similitud a obtener.

Teniendo en cuenta el alto poder de convocatoria de este evento y el nivel de los participantes, consideramos la tabla de resultados de esta tarea el mejor marco de referencia posible para validar nuestra propuesta. En la tabla 2 se muestran únicamente los mejores resultados de SemEval 2015.

Sistema	*Pearson*	Δ
Baseline-tokencos	0,529	0,0 %
RTM-DCU	0,582	5,4 %
UMDuluth	0,594	6,5 %
ExBThemis	**0,706**	**17,7 %**

Tabla 2: Resultados SemEval 2015

Valor	Frases
4	El espécimen es excepcional por las partes conservadas: un cráneo y mandíbula y un molde interno de la caja craneal. El espécimen comprende la mayor parte de la cara y mandíbula con los dientes y un molde interno de la caja craneal.
3	"Time 100" es una lista de las 100 personas más influyentes según la revista Time. La primera lista fue publicada en 1999 con las 100 personas más influyentes del siglo 20.
2	La "marinera" es un baile de pareja suelto, el más conocido de la costa del Perú. La marinera es el baile nacional del Perú, y su ejecución busca hacerse con derroche de gracia, picardía y destreza.
1	La "cripta de Santa Leocadia" está situada en el interior de la catedral de Oviedo, Asturias. Esteban Báthory fue sepultado en la cripta de la catedral de Wawel en Cracovia.
0	El río atraviesa la importante ciudad de Puebla de Zaragoza, la cuarta más poblada del país. El "Grêmio Esportivo Bagé" es un club de fútbol brasileño, de la ciudad de Bagé en el estado de Rio Grande do Sul.

Tabla 3: Ejemplos de pares de frases para cada nivel de similitud

4.1 Conjunto de datos

El conjunto de datos contiene pares de oraciones extraídas de los artículos de la Wikipedia y consta como es habitual de dos partes, una de *train* y otra de *test*. Cada parte tiene 324 y 251 pares de oraciones respectivamente. En la tabla 3 se muestra un ejemplo extraído del conjunto de datos para cada uno de los niveles de similitud establecidos para la tarea. Los valores de similitud de las oraciones fueron asignados calculando la media de las asignaciones manuales realizadas por cinco jueces humanos. Las oraciones en general están bien construidas y usan un lenguaje formal.

Se realizó un pequeño pre-procesado que consistió únicamente en eliminar los signos de puntuación. Aunque es una práctica común, la eliminación de palabras huecas o *stop words* no se llevó a cabo al considerarse que aportaban información relevante al proceso de alineamiento.

4.2 Modelos Word2Vec

El sistema de *word embedding* que hemos empleado en esta propuesta para obtener la representación vectorial de las palabras está basado en la implementación de Word2Vec (Mikolov et al., 2013) que encontramos en la herramienta Gensim (Řehůřek y Sojka, 2010). A través de ella hemos generado un primer modelo (Modelo-1) con textos en español provenientes de artículos de la Wikipedia[2], manteniendo así el mismo dominio que encontramos en los datos de la tarea de SemEval que nos sirve de referencia.

Este modelo se creó con vectores de 300 dimensiones, haciendo uso de la opción *negative sample* para eliminar palabras de ruido y ejecutando un total de 20 iteraciones para reforzar el entrenamiento y mejorar los resultados, siendo cada vez más estricto el factor de aprendizaje en cada una de estas etapas.

Además de generar este nuevo modelo también se han llevado a cabo experimentos utilizando un modelo extendido (Cardellino, 2016) (Modelo-2). Para su construcción se utilizaron, además de textos de la Wikipedia, otras fuentes de datos como el corpus AnCora-ES [3] o Europarl [4].

La experimentación con este modelo extendido nos permitirá comprobar los efectos de introducir en el modelo palabras extraídas de otros tipos de documentos con sus respectivos contextos, lo cual afectará a las distintas métricas aquí expuestas que hacen uso de dicho modelo.

4.3 Resultados

A la hora de evaluar los resultados, se ha utilizado la medida estadística que se aplicó en la tarea original del SemEval, el coeficiente de correlación de Pearson.

En las tablas 4 y 5 se muestran los resultados obtenidos por los métodos aquí propuestos, así como el mejor resultado registrado en la tarea de SemEval 2015 (ExBThemis). Aparecen tanto los métodos individuales como el resultado de combinar dichos métodos junto a la diferencia del resultado respecto al

[2]https://dumps.wikimedia.org/eswiki/latest/

[3]http://clic.ub.edu/corpus/

[4]http://www.statmt.org/europarl/

sistema ExBThemis. En la tabla 4 vemos los resultados obtenidos con el modelo de *word embeddings* creado exclusivamente con datos de la Wikipedia (Modelo-1), mientras que en la tabla 5 vemos los resultados obtenidos utilizando el modelo extendido con datos adicionales que no provienen de la Wikipedia (Modelo-2).

Sistema	*Pearson*	Δ
ExBThemis	0,706	-
Euclidea (E)	0,509	-19,7 %
Coseno (C)	0,467	-23,9 %
Alineamiento (A)	0,692	-1,4 %
Combinado (E+C+A)	**0,713**	**0,7 %**

Tabla 4: Resultados con el Modelo-1

Sistema	*Pearson*	Δ
ExBThemis	0,706	-
Euclidea (E)	0,642	-6,4 %
Coseno (C)	0,646	-5,9 %
Alineamiento (A)	0,687	-1,8 %
Combinado (E+C+A)	**0,723**	**1,8 %**

Tabla 5: Resultados con el Modelo-2

En ambos casos las métricas empleadas de forma individual no son capaces de obtener resultados que superen los obtenidos por el sistema ExBThemis, aunque el método de alineamiento aquí planteado se queda a menos de dos puntos porcentuales. Teniendo en cuenta que nos comparamos con el mejor sistema de entre todos los que se presentaron para resolver esta tarea en la edición 2015 de SemEval, dicho resultado puede considerarse relevante por sí mismo. Sin embargo, es la combinación de las diferentes métricas la que arroja las mejores cifras superando significativamente nuestra referencia, especialmente en el caso de utilizar el Modelo-2.

Precisamente en el mejor de los casos, utilizando el Modelo-2, la inserción de textos de diferentes dominios a la hora de generar el modelo de *word embedding* afecta negativamente al método de alineamiento, quizás por las diferencias en cuanto a las relaciones semánticas entre palabras que son aprendidas partiendo de dichos textos. Sin embargo, ese enriquecimiento del vocabulario favorece enormemente a las métricas más tradicionales, que consiguen resultados considerablemente mejores beneficiando con ello a la combinación, que compensa así la ligera pérdida sufrida por el alineamiento.

5 Conclusiones y trabajo futuro

En este trabajo hemos explorado la forma de aprovechar un modelo de *word embedding* para mejorar los resultados en una tarea de similitud semántica de textos. Nuestro principal objetivo era evaluar la mejora que se puede obtener en esta tarea sin hacer uso de otros recursos que no sean la representación vectorial en el espacio continuo. Para ello hemos definido dos maneras de calcular la similitud semántica de textos. Por un lado, mediante un alineamiento entre textos bidireccional y ponderado en función del parecido de las palabras emparejadas. La otra técnica se apoya en los vectores de palabras para construir representaciones vectoriales de los textos que son comparadas para determinar el grado de similitud entre ellos. Los experimentos sobre un corpus de la competición SemEval de 2015 para español muestran que el método de alineamiento se comporta de manera muy satisfactoria, quedando muy cerca del mejor sistema de la competición. En el caso de la técnica basada en la representación vectorial de textos, los resultados son peores pero aportan conocimiento complementario. Esto se demuestra con el hecho de que cuando se combinan las salidas de ambas técnicas se consiguen mejorar ambos resultados, superando incluso al mejor de los sistemas de la competición. También hemos experimentado con dos modelos distintos de *word embedding*, uno de ellos entrenado exclusivamente con textos de Wikipedia y otro, más extenso, también con textos de la Wikipedia más textos de otras fuentes. Dado que el corpus de SemEval está creado con textos de Wikipedia, este experimento nos ha permitido evaluar la sensibilidad de las técnicas a los textos usados para el entrenamiento de los modelos. Los experimentos han mostrado que la técnica basada en la representación vectorial de textos se beneficia de un modelo de *word embedding* más extenso aunque entrenado con textos de un dominio diferente. Por su parte, la técnica de alineamiento es penalizada levemente por el cambio de dominio aunque sus resultados siguen siendo mejores que los de la técnica de representación vectorial. En cualquier caso, la combinación sigue mejorando los resultados de ambas técnicas, quedando también por encima del mejor sistema de la competición. Como trabajo futuro estamos especialmente interesados en analizar el comportamiento de nuestros sistemas en un con-

texto multilingüe. El buen comportamiento de las técnicas de *word embedding* a la hora de comparar palabras de distintos idiomas (Mikolov, Le, y Sutskever, 2013) y el hecho de que nuestra aproximación sólo se base en la información de los modelos de *word embedding* nos hace pensar que puede comportarse bien para calcular la similitud de textos en distintos idiomas.

Agradecimientos

Este trabajo ha sido financiado a través del proyecto de investigación AORESCU (P11-TIC-7684 MO).

Bibliografía

Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, y R. Mihalcea. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. En *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, páginas 252–263.

Bär, D., T. Zesch, y I. Gurevych. 2013. Dkpro similarity: An open source framework for text similarity. En *ACL (Conference System Demonstrations)*, páginas 121–126.

Biçici, E. 2015. Rtm-dcu: Predicting semantic similarity with referential translation machines. *SemEval-2015*.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, y P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, Noviembre.

Hänig, C., R. Remus, y X. D. L. Puente. 2015. Exb themis: Extensive feature extraction from word alignments for semantic textual similarity. *SemEval-2015*, página 264.

Jiang, J. y D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Karumuri, S., V. Vuggumudi, y S. Chitirala. 2015. Umduluth-blueteam: Svcsts-a multilingual and chunk level semantic similarity system. *SemEval-2015*, página 107.

Lin, D. 1998. Extracting collocations from text corpora. En *First workshop on computational terminology*, páginas 57–63. Citeseer.

Mihalcea, R., C. Corley, y C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. En *AAAI*, volumen 6, páginas 775–780.

Mikolov, T., Q. Le, y I. Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, y J. Dean. 2013. Distributed representations of words and phrases and their compositionality. En *Advances in neural information processing systems*, páginas 3111–3119.

Miller, G. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Pennington, J., R. Socher, y C. Manning. 2014. Glove: Global vectors for word representation. En *EMNLP*, volumen 14, páginas 1532–1543.

Řehůřek, R. y P. Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. En *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, páginas 45–50, Valletta, Malta, Mayo. ELRA. `http://is.muni.cz/publication/884893/en`.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Zou, W., R. Socher, D. Cer, y C. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. En *EMNLP*, páginas 1393–1398.